Alexa at SemEval-2023 Task 10: Ensemble Modeling of DeBERTa and BERT Variations for Identifying Sexist Text

Mutaz Younes
Dep. of Computer Science
Maharishi International University
Iowa, USA
mutazyounes@gmail.com

Ali Kharabsheh
Dep. of Computer Science
Yarmouk University
Irbid, Jordan
alikharabsha12@gmail.com

Mohammad Bani Younes
Dep. of Computer Science
Ajloun National University
Irbid, Jordan
aliyounes1962@gmail.com

Abstract

This study presents an ensemble approach for detecting sexist text in the context of the Semeval-2023 task 10. Our approach leverages 18 models, including DeBERTa-v3-base models with different input sequence lengths, a BERT-based model trained on identifying hate speech, and three more models pre-trained on the task's unlabeled data with varying input lengths. The results of our framework on the development set show an f1-score of 84.92% and on the testing set 84.55%, effectively demonstrating the strength of the ensemble approach in getting accurate results.

1 Introduction

Sexist language has been a persistent issue in various forms of communication, including online communication. The use of sexist language perpetuates gender stereotypes and reinforces the marginalization of certain groups, particularly women. Detecting and addressing instances of sexist language is crucial for promoting gender equality and reducing discrimination.

In recent years, there has been a growing interest in developing methods for automatically detecting sexist language. The Semeval-2023 task 10 provides (Kirk et al., 2023) a platform for researchers to explore and evaluate various approaches to identifying sexist text. This task involves identifying instances of sexist language in various forms of online communication, including social media posts, comments, and reviews.

Previous research has explored different methods for identifying sexist language, including rule-based approaches, machine-learning techniques, and deep-learning models. While these approaches have shown promising results, the challenge of detecting sexist language remains complex due to the subtleties and nuances of language use.

Our approach for identifying the sexist language in the Semeval-2023 task 10 subtask A involves an ensemble of 18 different models, including DeBERTa-v3-base and pre-trained BERT-based models pre-trained on the task's unlabeled data. The model's predictions are combined to improve overall performance. Our method achieved an f1-score of 84.92% on the development dataset and 84.55% on the testing dataset, ranking 30th out of 89 participating teams.

Section 2 of this paper introduces an overview of the related work in the field of sexist text identification. Section 3 gives an overview of the details of the dataset we used in this research. Section 4 presents key characteristics of the ensemble approach and provides a detailed description of the models used. Finally, in Section 5, we conclude the paper and discuss future directions for research in this area.

2 Background

Sexism against women on social media has become a growing concern, leading researchers to develop automatic systems that detect sexist text. Online competitions such as the SemEval-2022 Task 5 on "Multimedia automatic misogyny identification" and "sEXism Identification in Social neTworks (EXIST)" at IberLEF 2021 (Fersini et al., 2022; Rodríguez-Sánchez et al., 2021) have helped accelerate this research. The competition aimed to identify sexism in social media content using machine learning, with two sub-tasks: binary classification and fine-grained classification distinguishing between five different types of sexist content.

19 teams participated in the EXIST 2021 benchmark and proposed different approaches to solve the task. The team UMUTeam (García-Díaz et al., 2021) combines linguistic features with state-of-the-art transformers to achieve an accuracy of 76.47% for binary classification (task 1) and ranks seventh. For the multi-class classification (task 2) they achieve an accuracy of 67.67% ranking third. Another team (Butt et al., 2021), presents

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1644–1649

their results on the shared task and emphasizes the importance of pre-processing techniques and data augmentation in overcoming the challenges posed by a multilingual dataset and inconsistencies of social media text. Their work achieves an F1 score of 78.02% for binary classification (task 1) ranking ninth and 49.08% for fine-grained classification (task 2) ranking fifth.

The AIT_FHSTP team (Mina et al., 2021) applies two multilingual transformer models, multilingual BERT and XLM-R, using pre-training with additional data and supervised fine-tuning with augmented data. The best model in their work is the XLM-R with a macro F1-score of 77.52% for binary classification (task 1) ranking 13 and 55.89% for the multi-class classification (task 2) ranking seventh.

(Kalra and Zubiaga, 2021) presents a deep neural network approach to detect sexist text, including BERT and DistilBERT. The best model for binary classification (task 1) uses BERT and a multi-filter CNN model achieving an accuracy of 76.2%. The same model with data augmentation achieves the best performance for multiclass classification with an F1 score of 51.9%

In earlier work, (Sharifirad et al., 2018) explore the use of knowledge graphs to improve the performance of sexist tweet classifiers. The authors propose using ConceptNet and Wikidata to improve sexist tweet classification by two methods: text augmentation and text generation. In the text generation approach, new tweets are generated by replacing words with data from ConceptNet relations, increasing the size of the training set and preserving the label. In the text augmentation approach, words in the tweets are augmented with words from ConceptNet relations and their descriptions from Wikidata, increasing the length of each tweet without changing the number of tweets in each class. The authors find that their approach significantly improved sexist tweet classification across multiple machine learning models and claim it can be applied to other small dataset problems, such as hate speech or abusive language and text classification.

3 DataSet

The Subtask A dataset used in this research consists of 20,000 entries sampled from Gab and Reddit. The training dataset, which constitutes 70% of the overall dataset, contains 14,000 entries, out of which 3,398 are labeled as sexist.

Dataset	Sentences	Percentage
Training	14000	70%
Testing	4000	10%
Development	2000	20%

Table 1: Number of Sentences in Training, Development, and Testing Datasets

To enable semi-supervised training techniques, the organizers also released two unlabeled datasets consisting of 1 million entries each from Gab and Reddit. The unlabeled datasets were prepared using the same cleaning and preparation procedures as the labeled dataset.

The development data constitutes 10% of the overall dataset and consists of 2,000 entries. The test data, which constitutes 20% of the overall dataset, contains 4,000 entries.

The number of sentences in the training and development datasets is summarized in Table 1:

4 System overview

Our system for identifying sexist language in the Semeval-2023 task 10 involves an ensemble approach that combines multiple transformers-based models (Vaswani et al., 2017), Section 4.2 provides details on these models. Specifically, we used a combination of different variations of DeBERTa-v3-base (He et al., 2020) and BERT (Devlin et al., 2018) models. The transformer models are fine-tuned on the provided labeled data and used to classify instances in the development and test sets. The code can be accessed at Code repository.

4.1 Data Preparation

Initially, we attempted to clean the training data by removing punctuations, converting text to lowercase, and removing extra spaces. However, this resulted in a decrease in performance on the development set, so we decided to keep the text as is. Additionally, we also explored the impact of data augmentation on the results. We tried various techniques such as using related data from previous tasks (Rodríguez-Sánchez et al., 2021) to augment the data. However, in general, data augmentation did not lead to a significant improvement in performance. The improvement in results came from ensembling the models rather than data augmentation.

4.2 Models Used

In this study, we explore different models to determine the optimal approach for a particular task. We utilize publicly available models and pre-train them on the unlabeled data provided by the task.

We explore a set of models such as RoBERTa models (Liu et al., 2019), BART models (Lewis et al., 2019), DeBERTa models (He et al., 2020), DistilBERT models (Sanh et al., 2019), and pre-trained BERT-based models. Based on the performance of each model on the development set, we eventually ended up choosing only deBERTa-v3-base and hateBERT (Caselli et al., 2020) and pre-trained them on the unlabeled data provided by the authors.

For both deBERTa-v3-base and hateBERT, we conducted a total of 18 fine-tuning experiments, each employing different combinations of learning rate and max length parameters. We explored various parameter values and ultimately selected those that yielded the highest scores, as shown in Table 2.

In our fine-tuning experiments with deBERTa-v3-base and hateBERT, the max length parameter proved to be an influential factor in determining the models' performance. The max length parameter specifies the maximum number of tokens the model processes in each input sequence. Adjusting this parameter allowed us to explore the trade-off between capturing sufficient context for accurate predictions and reducing the computational resources required.

Interestingly, our experiments revealed that sometimes shorter max length values resulted in better performance. This suggests that a more concise input sequence may provide the models with more focused and relevant context, enabling them to make more accurate predictions. On the other hand, longer max length values could introduce additional noise, potentially hindering the models' performance. We tested various max length values to identify the optimal setting for each model, as demonstrated in Table 2.

To provide a more detailed analysis, Table 2 displays the results of each model when tested individually on the development and test datasets. Several models, such as deBERTa-v3-base, achieve high f1 scores when used alone. Nevertheless, we observed a slight improvement in overall performance on the test set when combining all the predictions using a soft-ensemble approach, see section 4.3.

4.3 Ensemble Approach

We used a soft-ensemble method for combining the predictions of each model to produce the final prediction. The soft-ensemble method has been widely adopted in the field of natural language processing and enables the combination of probabilities, taking into account the confidence of each model's prediction. We also experiment with the hard ensemble method, which is another commonly used ensemble method. Hard ensemble involves selecting the most frequently predicted class from each model and using that as the final prediction. This method does not take into account the confidence of each model's prediction, therefore, we compared the performance of both soft-ensemble and hard-voting methods on the dataset to determine which method yielded the best results, see Table 3.

The approach we used involves each model producing a probability score whether it is sexist or not for each instance, and these scores are then combined using a weighted average to produce the final prediction. This method has been shown to be effective in improving the accuracy of predictions compared to using a single model (Risch and Krestel, 2020; Dang et al., 2020; Briskilal and Subalalitha, 2022).

Furthermore, we also explore the ability to eliminate some of the models randomly to increase the overall score. Our approach involves generating a random number and using it to select a corresponding number of models from a list of 18 models. We observed a slight improvement in performance when using this approach, but it takes time to see a slight improvement in the results because it is resource-intensive. Therefore, we did not report or submit the output of this method, but we will keep it for further research opportunities.

5 Results

The obtained results of the individual models and the ensemble approaches are presented in Table 2 and Table 3.

Our results indicate that ensembling multiple models can be an effective approach for identifying sexist text. The combination of different models provides a more robust solution, leveraging the strengths of each individual model. It is possible that adding even more models to the ensemble would lead to further improvements in performance.

During our error analysis, we found that our en-

Model	Max Length	Batch Size	Learning Rate	Dev	Test
BERT_racism3	80	32	3e-05	83.23%	83.03%
BERT_racism3	90	32	3e-05	82.74%	83.32%
BERT_racism3	100	32	3e-05	83.21%	82.64%
BERT_racism3	110	32	3e-05	83.28%	82.62%
BERT_racism4	70	32	3e-05	83.34%	83.33%
BERT_racism4	80	32	3e-05	82.77%	83.27%
BERT_racism4	90	32	3e-05	82.93%	83.29%
BERT_racism4	100	32	3e-05	83.29%	82.59%
BERT_racism4	110	32	3e-05	82.92%	82.58%
deBERTa-v3-base	50	32	5e-05	82.67%	83.37%
deBERTa-v3-base	60	32	5e-05	82.66%	83.17%
deBERTa-v3-base	70	32	5e-05	84.03%	83.96%
deBERTa-v3-base	80	32	5e-05	83.59%	84.01%
deBERTa-v3-base	90	32	5e-05	83.48%	84.44%
deBERTa-v3-base	100	32	5e-05	83.50%	82.87%
deBERTa-v3-base	110	32	5e-05	83.16%	83.28%
hateBERT	90	32	5e-05	83.98%	83.34%
BERT_hateracism90000	90	32	5e-05	82.64%	83.36%

Table 2: The 18 models used in our experiments. Results are based on the development data set. All models can be found at https://huggingface.co

Models	Dev	Test	Notes
Ensemble of 18 models	84.74%	84.46%	prediction with highest probability
Ensemble of 18 models	85.05%	84.45%	hard-ensemble method
Ensemble of 18 models	84.92%	**84.55%**	soft-ensemble method

Table 3: Results reported based on the development and test sets. The calculation of the prediction with the highest probability involves selecting the prediction from the most confident model, determined by the predicted probability of each label.

semble approach had difficulty identifying certain types of sexist sentences. Specifically, we struggled with detecting sexist language that involves immutable gender differences and gender stereotypes, as well as incitement and encouragement of harm. These types of sentences often involve subtle language use and required a deeper understanding of the context and underlying social issues.

Additionally, we found that our approach had difficulty with detecting dehumanizing attacks and overt sexual objectification. These types of language often involved explicit references to the target's body and required a fine-grained analysis of the language use.

6 Conclusion

In this research, we investigated the use of an ensemble approach for identifying sexist text. Our approach combined multiple transformer models, including different variations of DeBERTa-v3-base, and BERT. We found that the ensemble approach outperformed the individual models, achieving 84.92% f1-score on the development set of the Semeval task and 84.55% on the testing set.

Our results demonstrate the effectiveness of combining multiple models to identify sexist text. The combination of different models provides a more robust solution, leveraging the strengths of each individual model. Our approach also showed that pre-training some of the models on the unlabeled data provided by the task can be an effective way of incorporating relevant information.

Additionally, we explored the impact of data augmentation on the results. Our findings indicate that data augmentation techniques, such as using online data on the same topic, did not result in a significant improvement in performance. The improvement in results came from ensembling the

models rather than data augmentation.

Our conclusion highlights the importance of ensembling multiple models for the task of identifying sexist text. Further research is needed to determine the optimal number of models to include in an ensemble and to improve the performance of the ensemble approach.

We discovered that our system struggles with identifying certain forms of subtle and implicit sexism, which is a common challenge in detecting sexist language. In future work, we plan to explore additional feature engineering techniques and alternative methods for model selection to further improve the performance of our ensemble system.

Overall, this research provides valuable insights into the use of an ensemble approach for identifying sexist text. The findings have important implications for future work in this area and demonstrate the potential for using ensembles to tackle complex NLP tasks.

References

J Briskilal and CN Subalalitha. 2022. An ensemble model for classifying idioms and literal texts using bert and roberta. *Information Processing & Management*, 59(1):102756.

Sabur Butt, Noman Ashraf, Grigori Sidorov, and Alexander F Gelbukh. 2021. Sexism identification using bert and data augmentation-exist2021. In *IberLEF@ SEPLN*, pages 381–389.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2020. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Huong Dang, Kahyun Lee, Sam Henry, and Ozlem Uzuner. 2020. Ensemble bert for classifying medication-mentioning tweets. In *Proceedings of the Fifth Social Media Mining for Health Applications Workshop & Shared Task*, pages 37–41.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Elisabetta Fersini, Francesca Gasparini, Giulia Rizzi, Aurora Saibene, Berta Chulvi, Paolo Rosso, Alyssa Lees, and Jeffrey Sorensen. 2022. Semeval-2022 task 5: Multimedia automatic misogyny identification. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 533–549.

José Antonio García-Díaz, Ricardo Colomo-Palacios, and Rafael Valencia-García. 2021. Umuteam at exist 2021: Sexist language identification based on linguistic features and transformers in spanish and english.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Amikul Kalra and Arkaitz Zubiaga. 2021. Sexism identification in tweets and gabs using deep neural networks. *arXiv preprint arXiv:2111.03612*.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Schütz Mina, Boeck Jaqueline, Liakhovets Daria, Slijepčević Djordje, Kirchknopf Armin, Hecht Manuel, Bogensperger Johannes, Schlarb Sven, Schindler Alexander, and Zeppelzauer Matthias. 2021. Automatic sexism detection with multilingual transformer models. *arXiv preprint arXiv:2106.04908*.

Julian Risch and Ralf Krestel. 2020. Bagging bert models for robust aggression identification. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 55–61.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Procesamiento del Lenguaje Natural*, 67:195–207.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Sima Sharifirad, Borna Jafarpour, and Stan Matwin. 2018. Boosting text classification performance on sexist tweets by text augmentation and text generation using a combination of knowledge graphs. In *Proceedings of the 2nd workshop on abusive language online (ALW2)*, pages 107–114.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all

you need. *Advances in neural information processing systems*, 30.